ARABIC SIGNATURE VERIFICATION DATASET

SOMAYA AL-MAADEED, WAEL AYOUBY, ABDELAALI HASSAINE, AMNA ALMEJALI, AMAL AL-YAZEDI, and ROUDHA AL-ATIYA

Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

Abstract
In this paper two new signature verification datasets in the offline and on/off-line mode are presented. Then there are a full description and analysis of these datasets in different stages and in both modes. These datasets are mainly based on Arabic signatures. The signature verification concept requires a huge number of signatures, therefore the two datasets, in the offline and the on/off-line mode, separately collected, contain thousands of signatures from contributors belonging to different ages, nationalities, genders and academic levels. These signatures will be used for the training, testing and verification step. The offline dataset contains around 12000 signatures for more than 1000 volunteers and the on/off-line dataset contains more than 1750 reference and disguise signatures for more than 190 volunteers. Both datasets will be exceptional given the number of contributors as well as the number of signatures collected.

Keywords: signature verification, Arabic datasets, offline handwriting, and writer identification

1. INTRODUCTION

These days the signature verification concept is an authentication method recognized socially and used regularly to improve the identity of people, to do verifications on the confidential systems and to solve security problems caused by fake signatures that criminals adopt in their techniques. Being an exciting problem, the automatic signature verification by some software has received a large interest concerning pattern recognition in several universities and research centers. Researches concerning this concept are numerous, and a large number of datasets were collected to improve methods used by research teams, and different algorithms have been proposed by various researchers to solve problems caused by forged signatures.

2. EXISTING DATASETS

In the Chinese University of Hong Kong, a team has used a dataset in the online mode [1]. It is used in the Signature Verification Competition (SVC) 2004 [2] containing English and Chinese signatures, done on WACOM Intuos tablet. The dataset consists of 3200 signatures divided into two datasets. Each dataset contains signatures of 40 different writers. Each writer has written 20 genuine signatures, 10 signatures in a 1st session, and 10 others in a 2nd session. Volunteers also contributed in four skilled forgeries for each of five other contributors (40 subject each one containing 40 signatures: 20 genuine and 20 forgery signatures).

Also, a team in Michigan State University in USA used the online datasets DB1 and DB2 to evaluate their method [3]. DB1 contains 520 signatures for 52 writers, then for 20 writers of them, individuals were asked to produce 3 skilled forgeries (60 forgery signatures in total). Then the dataset DB2 which is a superset of DB1 contains 1232 signatures for 102 different writers, including forgeries classified into random and skilled forgeries [3].

Furthermore, concerning the Off-line datasets, researchers in Griffith University in Australia used, to evaluate their method and perform the experiments [4], the GPDS signature dataset [5] in the offline mode. Only 39 sets of signatures in this dataset were available at the time of the experimentation. They used a total of 2106 signatures to improve the results of the method. These 2106 signatures are composed by 39 sets of signatures for 39 persons. Each set contains 24 samples of genuine and 30 samples of forgeries.

Then a new dataset called "SigSA" [6] was collected in the online mode in Turkey, using Interlink Electronics ePad-ink tablet, with a pressure sensitive touchpad (300 dpi, 100Hz). This dataset contains signatures for 110 volunteers. Each signer was asked to do 20 samples of his signature in two different sessions separated by one week (1100 genuine signatures in the 1st session and 1100 other signatures in the 2nd session). Then five forgery signatures were obtained for each subject (550 forgeries in total).

As well, researchers in the Anna University Information Technology Chennai in INDIA collected their own dataset [7] in the offline mode to verify signatures on the checks. The documents were
digitized with 400 dpi in JPG format. The dataset consists of 20 signatures collected from each customer and stored in the dataset.

Additionally, other datasets in offline and online mode were collected [8] by the UNICAMP in Brazil and the “Universidad Tuiuti do Paraná, Curitiba, PR, Brazil”. The first dataset is an on/off-line dataset, composed by 400 signatures, collected by 20 different writers. Online dataset was collected by a digitizing tablet (100Hz, 1000 dpi). The writers have written on white paper on top of tablet, and then the papers were digitized to form the offline dataset. A second dataset only in the off-line mode contained 500 signatures scanned and digitized at 300 dpi for 50 other writers.

Besides, the “Escuela Superior de Ciencias Experimentales y Tecnología Universidad Rey Juan Carlos” created their own dataset in the offline mode to evaluate their system [9] containing around 112 signatures. The dataset consists of 28 signers; each one of them had done 4 signatures (3 original and 1 forged signatures).

In addition, a new dataset containing 622 genuine signatures from 69 human writers was collected by the Digital Media Lab, in Sharif University of Technology, for their new on-line handwritten signature verification system [10]. As well, they have used 1010 trained forgeries.

Moreover in Bell Laboratories, Holmdel, NJ, Dr. Vishyjit S. NALWA who worked on "Automatic On-Line Signature Verification" [11] tested his method with three datasets DB1, DB2 and DB3 collected by Bell Laboratories. The Dataset 1 (DB1) was collected in Bell Laboratories using LCD writing tablet with a tethered pen. They have collected 904 genuine signatures from 59 different persons in two sessions (10 genuine signatures per person in the first session and 5 or 10 more genuine signatures in the second session). Then they collected 325 forgery signatures by 32 participants. The dataset DB2 was collected by an NCR 5990 LCD writing tablet with a tethered pen. It is composed by 982 genuine signatures for around 102 different signers and 401 forgery skilled signatures. DB3 is collected on the same tablet of DB2, but there were 790 reference signatures and 424 forgery signatures collected from 43 different volunteers [11].

In the Department of Computer Engineering MIT COE in Pune, they collected an offline dataset containing 10 scanned images of signature per person used only for training [12]. Also the Signal Theory Group in the University of Vigo in Spain used two datasets for their system [13]. The first one is the MCYT and the second one is BIOSECURE-DS2. In the MCYT bimodal biometric dataset [14] [15], the signatures were acquired with a pen tablet “WACOM INTOUS A6 USB”. In a sub-corpus of 100 users, they used genuine signatures and skilled forgeries (50 signatures per user; 25 are genuine and 25 are skilled forgeries done by other contributors). The BIOSECURE-DS2 [14] [15] contains dynamic signatures collected in two sessions for 104 contributors (30 genuine and 20 forgery signatures per contributor).

In the next section we present the datasets for signature verification collected in Qatar University.

3. QU-PRIP DATASETS

In the Department of Computer Science in Qatar University, we collected two new datasets for the Arabic signature verification; the first one in the offline mode and the second one in the online and offline mode. It is obvious that there is no Arabic signature dataset for the signature verification concept, therefore these new datasets will be the first two datasets mainly based on Arabic signatures. The volunteers are belonging to Arabic and other nationalities. Note that the datasets will be partly available [16]. The QU-PRIP signature verification datasets can be used for future research related to the domain of signature verification, as the first datasets containing both Arabic and English signatures. In the coming sub-section, we present a full description including all details about the quality of images and an analysis of the offline signature verification dataset.

3.1. OFFLINE DATASET

This first dataset is in the offline mode containing reference signatures for more than 1017 volunteers using variety of pens, stored in folders (1017 folders). This offline dataset of signature verification is part of a huge dataset containing handwritings and signatures [17]. In each folder there are around twelve signatures per writer, in Arabic or in English depending on the writer's background. Then it contains around 12,000 signatures in total. We suggested that each volunteer tries to have the same signature in the 12 occurrences and to do the signatures in two or more different colors or pens. The dataset was scanned by a professional quality scanner "EPSON GT-S80" without compression with a high quality resolution of 600 dots per pixel to provide reliable information, as A4 color images in
the JPEG format. For each contributor, the 12 occurrences of signatures were done using two or more different pens providing more complexity to the system. Around 306 contributors are Qatari, while around 190 are Lebanese, 101 are Palestinian, 104 are Egyptian, 68 are Jordanian, etc... The percentages of volunteers per nationality are shown in the figure 1 below.

![Figure 1 - Percentages of contributors per nationality for the offline mode](image)

The dataset was 52% written by females and 48% written by males, who used right or left hand in writing. About 953 volunteers are right handed, while 64 volunteers are left handed. Around 530 female writers helped us with collecting forms, while the remaining 487 writers were males.

In the figure 2, there are some samples for offline Arabic signatures. These signatures will be segmented and separated each one in an image.

![Figure 2- Samples of off-line signatures](image)

This dataset for signature verification was collected from March till December 2011. Volunteers are from different age range. Some of them are older than 40 years, or between 30 and 39 years old, or between 18 and 29 years old, teenagers or even younger than 12 years old. Contributors are volunteers from Qatar University including students, professors and administrators, from the Qatar International School (QIS) and the Lebanese School in Doha-Qatar, also from the Lebanese University - Faculty of Engineering, some confidential companies and the “Balamand” School in Lebanon. For around 1017 contributors, each contributor’s folder containing 12 signatures (around 5-6 MB initially), the space required in total is around 6 GB.

There was a big interest in changing places and locations during the collection of this dataset, thus the volunteers were employees, university students, school students, workers, professors, engineers, etc. Therefore storing all details about contributors like nationality, position, sex, academic level and age in the system was a real necessity in providing analysis of the dataset, in case if an increase of the number of volunteers belonging to a specific category mentioned and stored in the dataset analysis file is required. In the next sub-section, the on/off-line dataset collected on a graphic tablet and containing thousands of on/off-line signatures is fully described.

3.2- ON/OFF-LINE DATASET

The second dataset is in the on/off-line mode and the signatures are in Arabic and English. In this new dataset the main focus is on the Arabic volunteers who have Arabic signatures. We used the Intuos 4 (large size) and Intuos 5 (medium size) touch professional graphic pen tablet with normal sensitivity. The features include position in axis (X-coordinate, Y-coordinate), Time stamp, Button status, Pressure, azimuth angle and altitude angle referring to the tablet. The screen dimensions of the Intuos 5 used are 223.5 mm (H) × 139.7 mm (V) (8.8 in × 5.5 in). The tablet is offering 5080 lines per inch resolution and there are 2048 levels of pressure in the Wacom pen. The sampling rate is 100 dots per second.

The online signatures are collected from 194 volunteers, using the same electronic inking pen. Six signatures as reference signatures have been done by the volunteer naturally, with normal dynamics and without trying to change his own signature, thus these signatures should be done in the same duration and the same manner (normally) on the graphic tablet. Then, and for each contributor, we also collect from the same volunteer three disguised signatures. The volunteer tries to do signatures similar to the reference but with slight modifications. The contributors are asked to do their own signatures slowly or to change the direction and the way of signing, to add or remove a character or something
else from the reference signature. In that way, the contributor of the signature will prove someday that the disguise signatures are not related to him, but to someone else who tried to imitate and forge his signature. Then the online registration in the files will mark some differences regarding the online features. The final shape of the disguise signatures should be close to the reference signature’s shape, but with slight differences. In addition to reference and disguise signatures, 50 subjects contain each one four forgeries. All of those fake forgery signatures weren’t written randomly but were skillful and trained forgeries; a high percentage of forgers have tried to create similar shape of reference signatures. Forgers were asked to imitate shape of the signatures. Then they have tried to do forgery signatures as close as possible to the genuine signatures taking into consideration the shape of the reference signatures. In the future, the forgeries will be collected for all remaining signatures in the dataset.

The online signatures are done with a Wacom inking pen on a white A4 sheet placed on the top of the graphic tablet. So for this online dataset, the signatures are also archived in hard copies and will be digitized to form the offline version. Then this dataset collected is stored in the online and the offline mode. Regarding the signatures, the difference should be matched not only through the online mode due to several features listed previously, but also through the offline mode. Even if the shapes of disguise signatures are similar to the references, some little changes are included as we insisted on volunteers. Totally the dataset contains around 1164 reference signatures collected by around 194 contributors, and 582 disguise signatures from the same persons.

Signatures have been collected in Qatar University in March and April 2012, and volunteers are students, master students, senior students, engineers, research assistants, Post-doc, teaching assistant, professors, coordinators, secretary and administrators. In our team, we have planned to collect more online signatures from different age ranges and nationalities, but the main focus is to collect Arabic signatures from Arabic volunteers. In another step we will collect forgery signatures from other authors. Then the contributors will be asked to try to imitate reference signatures after doing some training to have skilled forged signatures.

The signatures collected from the 194 contributors, who are almost from Arabic nationalities, are collected in Qatar University and other places in Qatar. In this on-off-line dataset, around 137 volunteers have contributed in the Arabic signatures. These 137 contributors are from Arabic nationalities; then there are around 57 non-Arabic signatures for the 57 remaining contributors who are belonging to Arabic and other nationalities and backgrounds (only 11 volunteers are non-Arabic).

In the figure 4, the nationalities of contributors helping in collecting the datasets are sorted from the highest to the smallest proportion. The volunteers whose nationalities are not attached (figure 4) are all Arabic people. The tables of analysis below show the number of volunteers per nationality (table 1) and age range (table 2).

![Reference Signatures](image1.png) ![Disguise Signatures](image2.png)

**Figure 3- Samples of on/off-line signatures**

In the figure 3 some samples of on-line signatures are shown, six are reference signatures and three are disguise signatures. These online signatures are registered in a folder for online signature verification, each reference signature is titled as [num_of_contributor][00][num_of_ref_sig], and each disguise signature is named as [num_of_contributor][d00][num_of_disguise] both having the “.gif” extension. For each of these signature images is associated an “.hwr” file containing the online features and dynamic information. Each file of these is titled by the same name of the signature with an “.hwr” at the end. The space required for these files is low since each file needs 10 kb maximum to be stored. Then for 194 contributors, a maximum of 30 MB is required, including all online signatures (references and disguise signatures). For the step of collecting the forged skilled signatures, other 10MB are required.
As for the offline signatures dataset, the diversity of volunteers in the on/off-line dataset is an important factor, and then there was a necessity, during the collection of signatures, to have volunteers belonging to different age range. The way of doing signatures differs between volunteers, especially volunteers from different age range. Accordingly, a student of 18 or 19 years old in the university does signatures differently than another volunteer who is working as an administrative assistant for long period, or a professor in the university or an engineer. The university student may take more time than another older volunteer, and may have different pressure on the graphic tablet than a contributor skilled in doing signatures. In this dataset females contributed as well as males, thus around 34% of volunteers are females (66 females) and 66% are males (128 men). Also the females may have less pressure on the tablet than males, and may have different way in signing especially in Arabic.

Therefore this on/off-line dataset is diverse and contains a mixture of nationalities, age ranges, sexes and positions. For this on/off-line dataset, an analysis file is stored, containing all details and information about the people who helped in collecting the signatures. This file is important for the same reason mentioned in the section of offline dataset. People younger than 15 years are not involved in our on/off-line dataset due to the lack of experience they have in doing signatures. In that way, they will not be able to do the same reference signatures several times without differences, or to have similarity in their normal signatures as much as required for the system criteria. The other lack also exists for the step of disguise signatures, since they should introduce slight differences which might be difficult for young students to do, especially because they should understand the goal of doing these signatures. Also for the step of forgery skilled signatures, these young people will not be able to do such a task at all, since it will need skills and intelligence to be done. Therefore in the step of forgery signatures, the people contributing in the task should be adults and would be trained to forge and imitate the reference signatures before collecting it.

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palestinian</td>
<td>32</td>
</tr>
<tr>
<td>Egyptian</td>
<td>28</td>
</tr>
<tr>
<td>Tunisian</td>
<td>18</td>
</tr>
<tr>
<td>Lebanese</td>
<td>11</td>
</tr>
<tr>
<td>Syrian</td>
<td>10</td>
</tr>
<tr>
<td>Qatari</td>
<td>8</td>
</tr>
<tr>
<td>Not Attached</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-19 years</td>
<td>6</td>
</tr>
<tr>
<td>20-24 years</td>
<td>71</td>
</tr>
<tr>
<td>25-29 years</td>
<td>47</td>
</tr>
<tr>
<td>30-34 years</td>
<td>24</td>
</tr>
<tr>
<td>older than 35 years</td>
<td>46</td>
</tr>
</tbody>
</table>

4. CONCLUSIONS

In this paper we presented a description of the new signature verification datasets, in on/off-line and
 offline modes, containing mostly Arabic signatures from Arabic people. This is the first in collection of offline and on-line signature datasets, including Arabic signatures. Both of datasets contain signatures from volunteers belonging to different nationalities. The importance of our datasets is due to the huge number of offline signatures (12,000 signatures) and on/off-line signatures (1800 signatures) from the large number of contributors. Due to this large number of signatures, the datasets would not be totally but partly available [16] for research projects and academic researchers in the future.

Acknowledgements

This work was supported by Qatar National Research Fund (QNRF) grant through National Priority Research Program (NPRP) No. 09 - 864 - 1 – 128 and Qatar University Student Grant No. QUST-CENG-FALL-11/12-10. The information and statements made in this paper are exclusively the responsibility of the authors and the project’s team members.

References:

[1] Yu Qiao, Jianzhuang Liu, Xiaou Tang "Offline Signature Verification Using Online Handwriting Registration”


[6] Alisher Kholmatov and Berrin Yanikoglu " SigSA: On-line Handwritten Signature Dataset”


[9] José F. Vélez, Ángel Sánchez and A. Belén Moreno "ROBUST OFF-LINE SIGNATURE VERIFICATION USING COMPRESSION NETWORKS AND POSITIONAL CUTTINGS"


